submitted to JASA
BU-1216-M-f.

Cross-validation, the Bootstrap, and Related Methods for Tuning Parameter Selection

Naomi Altman *
Biometrics Unit
Cornell University

Christian Léger †
Département de mathématiques
et de statistique
Université de Montréal

September 1994

Abstract

The dramatic increase in computing power available to statistical researchers has engendered a corresponding increase in computer intensive nonparametric estimation and prediction procedures. Many of these procedures are data-adaptive in the sense that the set of estimators (predictors) is in a large nonparametric class indexed by a tuning (smoothing) parameter, and that this tuning parameter is chosen using the data. This article discusses two methods of tuning parameter selection, cross-validation and the bootstrap, as well as a number of related methods based on the leave-some-out idea.

The methods are compared across a number of examples designed to illustrate implementation issues, the principles underlying selection techniques and consistency of the selected parameters. Cross-validation is readily implemented in prediction problems, but is not easily extended to problems which do not involve prediction, and may not lead to consistent parameter selection. The bootstrap is more difficult to implement in many situations of interest, but is consistent for some problems for which cross-validation and related methods are not. With some extra computational effort (but little extra human effort) the

*Supported by Hatch Grant 151410 NYF

†Supported by NSERC (Canada) and FCAR (Québec)

bootstrap technique can also provide confidence intervals for the estimator which takes into account the process of selecting the tuning parameter.

Keywords: Bandwidth selection, smoothing, modified maximum likelihood, kernel estimators, risk estimators.

1 Introduction

Many modern data modeling techniques are adaptive in the sense that, rather than depending on a parametric model, they involve only mild regularity conditions, such as smoothness, and a family of estimators indexed by a tuning or smoothing parameter that controls some trade-off between fidelity to the data and model complexity. Examples discussed here include the symmetric location problem using trimmed means, for which the tuning parameter is the trimming proportion, and kernel methods for regression, derivative, density and distribution function estimation, for which the tuning parameter is the bandwidth. These techniques are often used as estimation steps in more computer intensive methods such as generalized additive models (Hastie and Tibshirani, 1990), classification and regression trees (CART Breiman, et al, 1984) and multivariate adaptive regression splines (MARS, Friedman, 1991). Section 2 of this article introduces the examples that are discussed in detail.

Appropriate choice of tuning parameters is often critical for good performance of the resulting estimator. Optimal parameter choice is generally defined in terms of minimizing some measure of risk and the data-adaptive choice of smoothing parameter minimizes an estimator of risk. A number of risk estimators have been proposed. Prominent among these are cross-validation (CV) (Stone, 1974) and more recently the bootstrap (e.g., Léger and Romano, 1990a). A number of ad hoc methods are also in use. As estimation methodology becomes more complex, there is an increasing need to understand the process of adaptive selection of the tuning parameter. This paper explores some of the issues in tuning parameter selection, focusing on the use of CV and related estimators and bootstrap techniques.

Cross-validation estimates *prediction* risk by averaging losses when predicting the i^{th} observation by a leave-one-out predictor. For squared error loss, prediction risk and estimation risk are equivalent for the purpose of choosing a tuning parameter since they differ by a constant

independent of the tuning parameter. For other losses, the optimal estimator may differ from the optimal predictor. In Section 3 the consistency of squared error prediction risk estimators based on average prediction losses is shown to depend on the rate of convergence of the estimators: for estimators which converge at the rate $n^{1/2}$ they do not work, whereas they do work for estimators with slower rates of convergence such as nonparametric smoothers.

The term "cross-validation" is often used for leave-one-out risk estimators in contexts where prediction is not defined, such as in density estimation, distribution function estimation, or estimation of the derivative of a nonparametric regression curve. Examples and clarifications of what we call the leave-some-out principle are contained in Section 4.

Bootstrap risk estimators are computed by computing the risk for an estimated model. By a suitable modification of the method, the bootstrap can be used to estimate *either* prediction or estimation risk. Unlike estimators based on average prediction loss, the bootstrap can be used to select among $n^{1/2}$ convergent estimators. On the other hand we will see in Section 5 that bootstrap risk estimators are not necessarily simple to define. However because bootstrap methods estimate the distribution of the estimator, bootstrap methods can be used to construct confidence intervals as well as to choose the tuning parameter. This is an important advantage over other methods of choosing the tuning parameter.

Section 6 is a summary of our conclusions.

2 Examples of tuning parameter selection problems

The following examples encompass only a small fraction of the problems involving tuning parameter selection, but do illustrate many of the strengths and weaknesses of the methodologies discussed, and are the focus of much of the current work in this area. In each case, we specify the parameter of interest, θ or $\theta(x)$. The estimator will be called $\hat{\theta}_\lambda$. We begin with the simplest problem, the location problem.

Example 1: Trimmed means estimate of location

Let $y_1, \ldots, y_n$ be identically and independently distributed (i.i.d.) from a distribution symmetric about its median θ and let $y_{(1)}, \ldots, y_{(n)}$ be the order statistics. The (symmetric) λ-

trimmed mean is

$$\hat{\theta}_\lambda = \frac{1}{n - 2[n\lambda]} \sum_{i=[n\lambda+1]}^{n-[n\lambda]} y_{(i)}, \tag{1}$$

where $[\cdot]$ is the greatest integer function. The tuning parameter λ is the trimming proportion. Rosenberger and Gasko (1983) is a good introduction to trimmed means.

Example 2: Kernel nonparametric regression

Consider the regression problem:

$$y_i = \mu(x_i) + \epsilon_i, \qquad i = 1, \ldots, n, \tag{2}$$

where y_i is a dependent variable, x_i is a set of fixed design points, $\mu(x) = \theta(x) = E(y|x)$ is a smooth deterministic function, and ϵ_i are i.i.d. random variables with symmetric distribution and $E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma^2$.

Many kernel-based regression estimators have been introduced in the literature. Eubank (1988) and Härdle (1990) provide good introductions. We will use the Priestley-Chao kernel regression estimator (Priestley and Chao, 1972) which has the form

$$\hat{\theta}_\lambda(x) = \hat{\mu}_\lambda(x) = \frac{1}{n\lambda} \sum_{j=1}^{n} K\left(\frac{x - x_j}{\lambda}\right) y_j$$

where the kernel function, $K(x)$ is symmetric about 0 and integrates to 1 and the bandwidth, λ, is the tuning parameter.

Example 3: Kernel estimation of regression derivatives

Consider again the regression problem of Example 2. Often estimation of derivatives is of interest, for example, for growth rates and accelerations for human growth curves (Gasser and Müller, 1984). Then $\theta(x) = \mu^{(p)}(x)$. An estimator corresponding to the nonparametric regression estimator of Example 2 is

$$\hat{\theta}_\lambda(x) = \hat{\mu}_\lambda^{(p)}(x) = \frac{1}{n\lambda^{p+1}} \sum_{i=1}^{n} K^{(p)}\left(\frac{x - x_i}{\lambda}\right) y_i,$$

where $K^{(p)}(x)$ is the p^{th} derivative of a kernel K with bounded support, which is infinitely differentiable on its support and $p - 1$ times differentiable at the boundary. Note that $\hat{\mu}_\lambda^{(p)}(x)$ is just the p^{th} derivative of $\hat{\mu}_\lambda(x)$ defined in Example 2. A related estimator using somewhat different kernel weights was introduced by Gasser and Müller (1984).

Example 4: Kernel nonparametric density estimation

Let y_i be distributed i.i.d. F with a density $f(y)$. In this problem, θ is the density function, $\theta(y) = f(y)$. The kernel density estimator introduced by Rosenblatt (1956) has the form

$$\hat{\theta}_\lambda(y) = \hat{f}_\lambda(y) = \frac{1}{n\lambda}\sum_{i=1}^{n} K\left(\frac{y - y_i}{\lambda}\right)$$

where the kernel function has the same properties as in Example 2. For a good introduction to nonparametric density estimation, see Silverman (1986).

Example 5: Smooth estimation of the distribution function

Let $y_1, \ldots, y_n$ be i.i.d. from the distribution function F and density f. An estimator of the distribution function which is smoother than the empirical distribution function is the kernel distribution estimator of Nadaraya (1964)

$$\hat{\theta}_\lambda(y) = \hat{F}_\lambda(y) = \frac{1}{n}\sum_{i=1}^{n} H\left(\frac{y - y_i}{\lambda}\right), \tag{3}$$

where H is defined from a kernel K through $H(x) = \int_{-\infty}^{x} K(t)\,dt$. Here $\theta(y) = F(y)$ is the distribution function. Note that $\hat{F}_\lambda(y) = \int_{-\infty}^{y} \hat{f}_\lambda(t)\,dt$, where $\hat{f}_\lambda$ is the kernel density estimator of Example 4.

3 Cross-validation estimates of prediction risk

Cross-validation for selection of smoothing parameters has its origins in the validation of statistical prediction procedures by data splitting. Stone (1974) and Geisser (1975) recognized that CV can be used to choose among members of a family of prediction procedures by selecting the procedure which minimizes the CV estimator of prediction risk. The prediction risk of the predictor $\hat{y}_{\lambda,n}$ for predicting a new independent value of the process y^{new} at x_i is $R[y_i^{\text{new}}(x_i), \hat{y}_{\lambda,n}(x_i)]$. The CV estimator is defined by:

$$CV(\lambda, n) = 1/n\sum_{i=1}^{n} L[y_i, \hat{y}^{-i}_{\lambda,n-1}(x_i)], \tag{4}$$

where L is the loss function, $(y_1, x_1), \ldots, (y_n, x_n)$ are independent vectors of observations and $\hat{y}^{-i}_{\lambda,n-1}(x_i)$ is the predictor of y_i, indexed by λ, computed from all the data but y_i. (Some

problems, e.g., the location problem, do not have covariates x_i.) We then choose the predictor yielding the smallest value of $CV(\lambda, n)$.

Cross-validation is based on averaging prediction loss between the i^{th} observation y_i and its predictor based on the remaining observations, $\hat{y}^{-i}_{\lambda,n-1}(x_i)$. So $CV(\lambda, n)$ estimates the average *prediction* risk

$$1/n \sum_{i=1}^{n} R[y_i, \hat{y}_{\lambda,n-1}(x_i)], \tag{5}$$

where the risk R is the expected loss. If the objective of the statistical analysis is to find a ***predictor for future observations,*** minimizing prediction risk is clearly of interest, although we still need to check that $CV(\lambda)$ adequately approximates it. In many problems, however, the estimation risk, $R(\hat{\theta}_{\lambda,n}, \theta)$ is of more direct interest.

The values of λ which minimize $R[y_i^{\text{new}}, \hat{y}_{\lambda,n}(x_i)]$ and $R(\hat{\theta}_{\lambda,n}, \theta)$ need not be the same. But in the important special cases of location and regression estimation with squared error loss, $\hat{y}_{\lambda,n}(x_i) = \hat{\theta}_{\lambda,n}(x_i)$ and

$$\begin{aligned} R_2[y^{\text{new}}, \hat{\theta}_\lambda] &= E[y^{\text{new}} - \hat{\theta}_\lambda]^2 = E[y^{\text{new}} - \theta]^2 + E[\hat{\theta}_\lambda - \theta]^2 \\ &= Var[y^{\text{new}}] + R_2[\hat{\theta}_\lambda, \theta], \end{aligned}$$

since y^{new} and $\hat{\theta}_\lambda$ are independent. So prediction risk and estimation risk are equivalent for tuning parameter selection with squared error loss. Hence even if the goal is to minimize estimation risk, prediction risk estimators are often used.

However, even when prediction and estimation risk have minima at the same value of the tuning parameter, it is still necessary to check that the minimizer of average loss adequately estimates the minimizer of the risk. Consider the following decomposition of prediction loss:

$$\begin{aligned} L_2(y^{\text{new}}, \hat{\theta}_\lambda) &= (y^{\text{new}} - \hat{\theta}_\lambda)^2 \\ &= (y^{\text{new}} - \theta)^2 + 2(y^{\text{new}} - \theta)(\theta - \hat{\theta}_\lambda) + L_2(\theta, \hat{\theta}_\lambda). \end{aligned} \tag{6}$$

The cross-product term in this expansion has expectation zero, but for reasonable estimators $\hat{\theta}_\lambda$, the term of interest $R_2(\theta, \hat{\theta}_\lambda)$ converges to zero with n.

The cross-validation estimator of prediction risk (4) has a similar decomposition:

$$CV(\lambda, n) = 1/n \sum_{i=1}^{n} (y_i - \theta)^2 + 2/n \sum_{i=1}^{n} (y_i - \theta)(\theta - \hat{y}^{-i}_{\lambda,n-1}(x_i)) + 1/n \sum_{i=1}^{n} (\theta - \hat{y}^{-i}_{\lambda,n-1}(x_i))^2. \tag{7}$$

The three terms in (7) estimate the expectation of the corresponding terms in (6). In particular, the mean zero cross-product term is estimated by an average of (dependent) random variables with mean 0, and the last term approximates estimation risk which is also converging to zero. Hence cross-validation can only be successful if the cross-product term of CV converges to 0 *faster* than the estimation risk. This is the case in Example 2, but not in Example 1.

Example 2 (continued)

In nonparametric regression least squares CV was first introduced to choose the smoothing parameter in cubic smoothing splines by Wahba and Wold (1975) in the statistics literature and by Clark (1975) in an archeology journal. Under appropriate conditions, $n^{2/5}[\hat{y}_\lambda(x) - E(\hat{y}_\lambda(x))]$ converges to a normal distribution so that the estimation risk is $O(n^{-4/5})$ which is slower than the location problem where the estimation risk is typically $O(n^{-1})$. Härdle and Marron (1985) showed that $CV(\lambda)$ is asymptotically loss optimal for (weighted) least squares loss in the sense that if $\hat{\lambda}$ is the value of λ which minimizes the CV criterion

$$\lim_n \to \infty \left[\frac{d(\hat{\theta}_{\hat{\lambda}}, \theta)}{\inf_{\lambda \in H_n} L(\hat{\theta}_\lambda, \theta)} \right] = 1,$$

in probability or with probability one. Härdle, Hall, and Marron (1988) showed that the selected bandwidth converges to the optimal bandwidth, but the relative rate of convergence of the cross-validation bandwidth to the best bandwidth is extremely slow: $(\hat{\lambda}_n^{\mathrm{CV}} - \lambda_n^{\mathrm{opt}})/\lambda_n^{\mathrm{opt}}$ converges to 0 at the rate $n^{-1/10}$, where λ_n^{opt} is the bandwidth minimizing $R(\hat{\theta}_\lambda, \theta)$.

Example 1 (continued)

The estimation risk of most location estimators converges to 0 at rate $O(n^{-1})$ which turns out to lead to non-optimality of CV for choosing between location estimators. Stone (1977) showed that CV cannot choose between the mean and the median and that under normality and squared error loss, CV will asymptotically choose the mean and the median with probabilities 0.4992 and 0.5008, respectively. This adaptive estimator has an efficiency of 0.818 compared to the mean. Similar results hold if absolute error loss (L_1) is used instead. Pruitt (1988) generalized this result to show that asymptotically cross-validation does not choose the best trimming proportion of an adaptive trimmed mean.

Altman and Léger (1994a) generalize these results by showing that for a family of location estimators that satisfy a weak differentiability condition, leave-one-out CV is not asymptotically

loss optimal and does not asymptotically choose the estimator with smallest risk (i.e. is not asymptotically risk optimal).

Theorem 1 *(Altman and Léger, 1994a) Suppose $y_1, \ldots, y_n$ is an i.i.d. sample from distribution F with mean θ and finite variance σ_y^2. Suppose $\hat{\theta}_1(\mathbf{y})$ and $\hat{\theta}_2(\mathbf{y})$ are both estimators of θ such that*

$$\hat{\theta}_j = \hat{\tau}_j + R_{n,j}$$

where $\hat{\tau}_j = \frac{1}{n}\sum_{i=1}^n h_j(y_i)$, $E[h_j(y_i)] = \theta$, $Var[h_j(y_i)] = \sigma_j^2 < \infty$. Assume that $E(R_{n,j}^2) = o(1/n)$, and that the asymptotic correlation between them is not ± 1. Then for any fixed d, leave-d-out CV is not asymptotically risk optimal for choosing between $\hat{\theta}_1(\mathbf{y})$ and $\hat{\theta}_2(\mathbf{y})$.

That the problem is related to the use of average prediction losses and the rate of convergence of the location estimators rather than the sample reuse inherent in cross-validation is shown by the following estimator of prediction risk. Suppose that $y_1, \ldots, y_n$ and $y_1^{\text{new}}, \ldots, y_n^{\text{new}}$ are two independent samples of i.i.d. random variables from distribution F. Let $\hat{\theta}_1(\mathbf{y})$ and $\hat{\theta}_2(\mathbf{y})$ be two location estimators constructed from $y_1, \ldots, y_n$. Consider the following estimator of prediction risk:

$$APL[\hat{\theta}(\mathbf{y})] = \frac{1}{n}\sum_{i=1}^n [y_i^* - \hat{\theta}(\mathbf{y})]^2 \tag{8}$$

$$= \frac{1}{n}\sum_{i=1}^n [y_i^* - \theta]^2 - 2\frac{1}{n}\sum_{i=1}^n [y_i^* - \theta][\hat{\theta}(\mathbf{y}) - \theta] + [\hat{\theta}(\mathbf{y}) - \theta]^2. \tag{9}$$

Here, $y_1^*, \ldots, y_n^*$ can be thought of as a validation sample and is usually not available in practice so that APL is not a real competitor to CV. The following theorem relates the asymptotic risk optimality of APL for choosing between $\hat{\theta}_1$ and $\hat{\theta}_2$ to the rate of convergence of the estimators. In particular, if they converge slowly (rate less than $\sqrt{n}$) average prediction loss is asymptotically loss optimal, but if they converge rapidly it is not even risk optimal.

Theorem 2 *(Altman and Léger, 1994a) Suppose $y_1, \ldots, y_n, y_1^*, \ldots, y_n^*$ is an i.i.d. sample from distribution F with mean θ and finite variance σ_y^2. Suppose $\hat{\theta}_1(\mathbf{y})$ and $\hat{\theta}_2(\mathbf{y})$ are both estimators of θ such that*

$$n^p \begin{pmatrix} \hat{\theta}_1(\mathbf{y}) - \theta \\ \hat{\theta}_2(\mathbf{y}) - \theta \end{pmatrix} \xrightarrow{D} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2 \end{bmatrix} \right)$$

where σ_i^2 is the asymptotic variance of $\hat{\theta}_i$ and ρ is the asymptotic correlation between the estimators. Assume also that $|\rho| \neq 1$. Then average prediction loss is asymptotically loss optimal if and only if $p < 1/2$. If $p = 1/2$ then average prediction loss is not risk optimal.

The problem is that the cross-product in (9) and the estimation loss (third term) are stochastically as large. Other similar methods, such as adjusted residual sums of squares, suffer the same problem (Altman and Léger, 1994a).

One solution is to use a larger validation sample size. In the context of cross-validation, the size of the validation sample can be increased by leaving out more than one observation when computing the estimator thus keeping more observations to validate this estimate. Specifically, for a fixed n, let $d = d_n$ be an integer less than n and $r = n - d$. Following Shao and Wu (1989), define $S_{n,r}$ to be the collection of subsets of $\{1, \ldots, n\}$ which have size r. For any $S = \{i_1, \ldots, i_r\} \in S_{n,r}$, let $\hat{\theta}^S = \hat{\theta}(y_{i_1}, \ldots, y_{i_r})$. The leave-$d$-out CV estimator of risk is

$$CV_d(\hat{\theta}_j) = \frac{1}{dN} \sum_S \left(\sum_{i \in S^C} (y_i - \hat{\theta}_j^S)^2 \right),$$

where S^C is the complement of the set S and $N = \binom{n}{r}$ is the number of subsets of size r. Under the assumptions of Theorem 1, Altman and Léger (1994a) show that leave-d-out CV is asymptotically risk optimal for choosing between $\hat{\theta}_1(\mathbf{y})$ and $\hat{\theta}_2(\mathbf{y})$, provided that $d/n \to 1$, and $r \to \infty$. Hence, most observations must be used for validation, but the number of observations used to estimate must go to infinity. A similar result for variable selection in multiple regression was obtained by Shao (1993).

4 Leave-some-out principle

CV is among the statistical estimation techniques that are based on repeatedly dropping one or more observations out of the data. Our goal in this section is to explain how "leave-one-out" techniques work. We demonstrate that "leave-one-out" estimators that work perfectly well in one context may not carry over very well to related problems.

When the focus of the statistical analysis is on estimation rather than prediction, tuning parameter selection must be based on an estimation risk. Often a risk estimator can be based

on a loss of the form $L(\hat{\theta}_\lambda, \tilde{\theta})$ where $\tilde{\theta}$ is a proxy for θ. For example, under squared error loss:

$$E\,L(\hat{\theta}_\lambda, \tilde{\theta}) = R(\hat{\theta}_\lambda, \theta) + E(\tilde{\theta} - \theta)^2 - 2E(\hat{\theta}_\lambda - \theta)(\tilde{\theta} - \theta)$$

Hence the expected value of this criterion is the estimation risk plus the mean squared error of $\tilde{\theta}$, which does not depend on the tuning parameter λ, plus a covariance term which depends on λ. The criterion is unlikely to be useful for choosing the tuning parameter unless the covariance term is smaller than the risk. One way to ensure this is to compute $\hat{\theta}_\lambda$ and $\tilde{\theta}$ from independent subsets of the full sample. This suggests partitioning the data into disjoint subsets, computing $\hat{\theta}_\lambda$ and $\tilde{\theta}$ on the two subsets and approximating the expectation by averaging the squared differences over different partitions. For the location problem and regression problems when y_i is used as a proxy for $\theta(x_i)$ this is identical to CV. Hence these methods are often referred to as CV. We prefer to reserve the phrase "cross-validation" for prediction based estimators of risk and refer to the other methods as leave-some-out estimators of risk, as the heuristic involved is not model validation.

Below is an example of the correct use of the leave-some out principle.

Example 3, continued

Müller, Stadtmüller, and Schmitt (1987) explored the problem of bandwidth choice for estimating regression derivatives under squared error estimation loss

$$R_2(\hat{\mu}^{(p)}, \mu_\lambda^{(p)}) = \int (\hat{\mu}^{(p)}(x) - \mu_\lambda^{(p)}(x))^2 dx.$$

They proposed the following leave-some-out estimator of this risk:

$$\sum_{i=1}^{n} [D_i^{(p)} - \hat{\mu}_{\lambda;-I}^{(p)}(x_i)]^2 \tag{10}$$

where $D_i^{(p)}$ is the finite difference defined iteratively by:

$$x_i^{(0)} = x_i,\ D_i^{(0)} = y_i, \qquad x_i^{(k)} = \frac{x_{i+1}^{(k-1)} + x_i^{(k-1)}}{2},\ D_i^{(k)} = \frac{D_{i+1}^{(k-1)} - D_i^{(k-1)}}{x_{i+1}^{(k-1)} - x_i^{(k-1)}},$$

and $\hat{\mu}_{\lambda;-i}(p)$ is defined like $\hat{\mu}_\lambda^{(p)}$ except that the $p+1$ observations in the set I used in defining D_i are not used.

Simple algebra shows that if $\hat{\mu}_\lambda^{(p)}$ rather than $\hat{\mu}_{\lambda;-I}^{(p)}$ was used in (10), the criterion would be biased by a term of the form $\kappa_i^{(p)}\sigma^2$ where $\kappa_i^{(p)}$ is defined by replacing y_j by $\frac{1}{n\lambda^{p+1}}K^{(p)}(\frac{x_i-x_j}{\lambda})$ in the expression for D_i. So the leave-some-out principle justifies criterion (10).

Unfortunately, the leave-some-out principle does not seem to be well understood, and is often applied incorrectly. In the following example, the proposed leave-one-out estimator does not sufficiently reduce the bias of the risk estimator, and in fact is asymptotically equivalent to a leave-none-out criterion. Neither criterion appears to be very good in practice.

Example 5, continued

Sarda (1993) suggested a leave-one-out procedure for selection of smoothing parameter for kernel distribution function estimation with risk

$$R[\hat{F}_\lambda(x), F(x)] = E\int[\hat{F}_\lambda(x) - F(x)]^2 w(x).$$

and risk estimator

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}[\hat{F}_\lambda^{-i}(x_i) - F_n(x_i)]^2 w(x_i),$$

where $\hat{F}_\lambda^{-i}(x_i) = \frac{1}{n-1}\sum_{i\neq j} H\left(\frac{x_i-x_j}{\lambda}\right)$ is the leave-one-out kernel distribution function estimator and F_n is the empirical distribution function. Altman and Léger (1994b) introduced the leave-none-out criterion

$$LNO(\lambda) = \frac{1}{n}\sum_{i=1}^{n}[\hat{F}_\lambda(x_i) - F_n(x_i)]^2 w(x_i)$$

and showed that the two are asymptotically equivalent. Moreover, they showed that the expected value of the derivative of either criterion is asymptotically positive which implies that asymptotically, the selected bandwidth will be the smallest available bandwidth. Simulations have confirmed this point.

The problem is that $\hat{F}_\lambda$ and F_n are both computed using the full dataset and so are correlated. Removing the i^{th} observation in the computation of $\hat{F}_\lambda$ only removes one of the numerous covariance terms. What could be done instead is to compute $\hat{F}_\lambda$ and F_n on separate subsets. How large should the two subsets be? Since the optimal bandwidth decreases with n, most observations should be used in $\hat{F}_\lambda$ unless an asymptotic adjustment can be made to relate a choice of bandwidth for $r << n$ to one for n. But that leaves few observations for F_n which is

then a poorer proxy for F. Altman and Léger (1994b) suggest instead directly estimating the asymptotically optimal bandwidth.

Another use of the leave-some-out idea in risk estimation is to adjust a naive estimator so that it has the correct expectation. These *ad hoc* methods depend on the problem and may not carry over simply even to related problems. One example is bandwidth selection for nonparametric density estimation.

Example 4, continued

A criterion often used to evaluate nonparametric density estimators is the integrated squared error risk:

$$\begin{aligned} R[\hat{f}_\lambda(x), f(x)] &= E\int(\hat{f}_\lambda(x) - f(x))^2\,dx \\ &= E\int \hat{f}_\lambda^2(x)\,dx - 2E\int \hat{f}_\lambda(x)f(x)\,dx + \int f^2(x)\,dx. \end{aligned} \tag{11}$$

Unlike the empirical distribution function of Example 3, no clear proxy exists for f in this problem. However, the third term in (11) does not depend on λ, so minimizing the first two terms is sufficient. Note that this decomposition depends on the L_2 loss and that it would not be possible if an L_1 loss function was used. The first term is easily estimated by $\int \hat{f}_\lambda^2(x)\,dx$. Since $\int \hat{f}_\lambda(x)f(x)\,dx = E\hat{f}_\lambda(Y)$ where the density of Y is independent of the observations used in the calculation of $\hat{f}_\lambda$ but comes from the same distribution, we have

$$E\left[\frac{1}{n}\sum_{i=1}^{n}\hat{f}_\lambda^{-i}(y_i)\right] = E\left[\int \hat{f}_\lambda(x)f(x)dx\right], \tag{12}$$

where $f_\lambda^{-i}(y_i) = \frac{1}{(n-1)\lambda}\sum_{j\neq i} K\left(\frac{x_i - x_j}{\lambda}\right)$. This leads to the criterion

$$LOO(\lambda) = \int \hat{f}_\lambda(x)^2 dx - \frac{2}{n}\sum_{i=1}^{n}\hat{f}_\lambda^{-i}(y_i)$$

which was introduced by Rudemo (1982) and also discussed by Hall (1983) and Stone (1984) who showed the asymptotic optimality of $LOO(\lambda)$. Rudemo (1982) and Bowman (1984) also introduced the "least squares cross-validation" criterion:

$$LSCV(\lambda) = 1/n\sum\int[\hat{f}_\lambda^{-i}(x)]^2 dx - \frac{2}{n}\sum_{i=1}^{n}\hat{f}_\lambda^{-i}(y_i) \tag{13}$$

using the predictive idea of cross-validation, but applied to *ad hoc* loss functions and *ad hoc* quantities to be predicted.

The criteria LOO and $LSCV$ are asymptotically equivalent so that the only place where a leave-some-out estimator must be used is in the cross-product term. If no observation is left out, then a bias term depending on λ as large as the risk is introduced.

The use of a leave-one-out estimator to estimate an expectation by a sum also plays a role in the modified maximum likelihood criterion introduced by Habbema, Hermans and van den Broek (1974) and Duin (1976) for bandwidth selection in density estimation. They defined the likelihood as the product of the density estimators evaluated at each observation, with λ in place of the parameter. Noting that this leads to a choice of $\lambda = 0$, they defined a modified likelihood as the product of the leave-one-out density estimators. This can be converted to the problem of minimizing Kullback-Leibler risk, by taking the logarithm of the likelihood. The leave-one-out term is then $\frac{1}{n}\sum_{i=1}^{n} log[\hat{f}_\lambda^{-i}(y_i)]$ which is an unbiased estimator of $E(\log[\hat{f}_\lambda(y_i)])$.

In the density estimation problem with squared error and Kullback-Leibler losses, the term in the risk involving the unknown parameter could easily be written as expectations of leave-one-out estimators. However, this does not extend to the distribution function estimation problem (Example 5) for which the corresponding term is $\int \hat{F}_\lambda(x)F(x)\,dx$ which is not the expectation of a quantity that can easily be approximated by a sample mean.

Modified maximum likelihood has also been suggested for bandwidth selection in non-parametric regression with binary response (Azzalini, Bowman and Härdle, 1989; Hastie and Tibshirani, 1990, p. 159 call it cross-validated deviance). However, the associated loss function is then

$$L(\mu, \hat{\mu}_\lambda) = \int \mu^2(x) log[\mu(x)/\hat{\mu}_\lambda(x)]dx + \int [1-\mu(x)]^2 log([1-\mu(x)]/[1-\hat{\mu}_\lambda(x)])dx$$

which is clearly not Kullback-Leibler loss. There do not appear to be any compelling reasons to consider this as an appropriate loss function for this problem. However, ordinary least squares CV is consistent for L_2 risk for binary regression (Altman and MacGibbon, 1993).

5 Bootstrap estimators of risk

The bootstrap, introduced by Efron (1979), is a method to estimate the sampling distributions, and can therefore be used to estimate risk. Léger, Politis and Romano (1992) is a recent survey of bootstrap techniques and Efron and Tibshirani (1993) is an excellent introduction. Early work on use of the bootstrap to choose tuning parameters include Härdle and Bowman (1988) who studied bandwidth selection in nonparametric regression and Hall and Martin (1988) who studied selection of a shrinkage parameter. Léger and Romano (1990a) studied bootstrap choice of tuning parameters in a general framework and obtained consistency and weak convergence results for a number of examples.

To illustrate, consider bootstrap estimators of estimation risk when θ is a scalar location parameter. Let $y_1, \ldots, y_n$ be i.i.d. F. In many instances, the parameter θ can be written as a functional of F, i.e., $\theta = \theta(F)$. The estimation risk of $\hat{\theta}_\lambda$ in estimating θ is

$$R[\theta(F), \hat{\theta}_\lambda] = E_F(L[\theta(F), \hat{\theta}_\lambda(y_1, \ldots, y_n)]), \tag{14}$$

where we have explicitly shown that the expectation is taken with respect to the random variables $y_1, \ldots, y_n$ with distribution F. If F is known, then the risk can either be computed explicitly, or simulated as follows.

In practice F is unknown. The bootstrap estimator of risk is computed by replacing F by an estimator $\hat{F}$, i.e., using $R(\theta(\hat{F}), \hat{\theta}_\lambda)$ as the estimator of $R(\theta(F), \hat{\theta}_\lambda)$. In most cases, one must resort to simulation. First, one computes $\theta(\hat{F})$. Then one generates a bootstrap sample $y_1^*, \ldots, y_n^*$ i.i.d. from $\hat{F}$ and computes $\hat{\theta}_\lambda(y_1^*, \ldots, y_n^*)$ and the bootstrap loss $L(\theta(\hat{F}), \hat{\theta}_\lambda(y_1^*, \ldots, y_n^*))$. Repeating a large number of times, the bootstrap estimator of risk is the average of the bootstrap losses.

The F may be estimated nonparametrically or parametrically. The most common nonparametric estimator of F is the empirical distribution function, leading to the usual resampling with replacement from the data. A smoother nonparametric estimator of F is the kernel estimator $\hat{F}_\lambda$ of (3). If it is assumed that F is in a parametric family F_β, then a parametric bootstrap would resample from $F_{\hat{\beta}}$ where $\hat{\beta}$ is a suitable estimator of β based on the original observations $y_1, \ldots, y_n$, such as a maximum likelihood estimator.

For the purpose of tuning parameter selection, we choose $\hat{\lambda}_n^{\text{Boot,est}}$ to be the value of λ yielding the smallest value of the bootstrap estimator of estimation risk $R(\theta(\hat{F}), \hat{\theta}_\lambda)$, i.e.,

$$\hat{\lambda}_n^{\text{Boot,est}} = \arg\min_{\lambda \in \Lambda} \mathrm{R}(\theta(\hat{\mathrm{F}}), \hat{\theta}_\lambda). \tag{15}$$

Of course, each estimator $\hat{F}$ defines its own bootstrap estimator $\hat{\lambda}_n^{\text{Boot,est}}$.

The bootstrap can also be used to estimate prediction risk. Let $y_1^*, \ldots, y_n^*, y^{\text{new},*}$ be i.i.d. from $\hat{F}$. Note that along with the bootstrapped version of the original sample, there is also a bootstrap independent future observation $y^{\text{new},*}$. Then one computes the bootstrap predictor $\hat{y}_\lambda^{\text{new}}(y_1^*, \ldots, y_n^*)$ and the bootstrap loss $L(y^{\text{new},*}, \hat{y}_\lambda^{\text{new}}(y_1^*, \ldots, y_n^*))$. Repeating a large number of times, the bootstrap estimator of prediction risk is the average of these bootstrap prediction losses. The value of λ corresponding to the smallest bootstrap estimator of prediction risk will be denoted by $\hat{\lambda}_n^{\text{Boot,pred}}$.

We immediately see an important advantage of the bootstrap over CV: by modifying the bootstrap algorithm accordingly, it can estimate prediction or estimation risk as desired, whereas cross-validation can only estimate prediction risk.

Example 1 (continued)

Léger and Romano (1990b) used the bootstrap to choose the trimming proportion of an adaptive trimmed mean. For each $\lambda \in \Lambda$, a bootstrap estimator of variance of the λ-trimmed mean $\hat{\theta}_\lambda$ is computed and the trimming proportion corresponding to the smallest estimate of variance, $\hat{\lambda}_n^{\text{Boot,var}}$, is selected, leading to the adaptive estimator of location $\hat{\theta}_{\hat{\lambda}_n^{\text{Boot,var}}}$. Under regularity conditions, the asymptotic variance of $\hat{\theta}_{\hat{\lambda}_n^{\text{Boot,var}}}$ is identical to that of the trimmed mean with smallest asymptotic variance. Hence, $\hat{\theta}_{\hat{\lambda}_n^{\text{Boot,var}}}$ is asymptotically risk optimal.

The bootstrap can also be used to estimate prediction risk rather than estimation risk. The cross-product term of the bootstrap prediction risk estimator is identically 0 in the bootstrap estimator because the estimator is a bona fide expectation, albeit from $\hat{F}_n$ rather than F.

In general, bootstrap risk estimation is more complex because of the need to generate the bootstrap samples. For instance, in the nonparametric regression problem $y_i = \theta(x_i) + \epsilon_i$ where $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. F. Thus we need a pilot estimator of $\theta(x_i)$, the parameter of interest, as well as of the unknown distribution.

Bootstrap methods are obtained by replacing unknown elements by estimates and usually require simulation to compute the estimated risk. Different estimators of the unknowns lead to different bootstrap estimators of risk. The bootstrap can also be used to estimate prediction risk, when appropriate. We illustrate the ideas in the estimation of a nonparametric regression function and its derivative, and in nonparametric density estimation. In these complex problems, the application of the bootstrap algorithm can be complicated.

Example 2 (continued)

In the nonparametric regression problem, the unknown distribution generates the errors. However, the errors are not observable, so estimates of the error distribution F must be based on residuals.

Let $\hat{\mu}_\alpha(x_i)$ be a preliminary estimator of $\mu(x_i)$ using the pilot bandwidth α and let $\epsilon_i(\alpha) = y_i - \hat{\mu}_\alpha(x_i)$ be the i^{th} residual, where we have explicitly shown the dependence of the residual on the bandwidth of the pilot nonparametric regression estimator. An estimator of F is $\hat{F}_n(\alpha)$, the empirical distribution function of the *centered* residuals, $\epsilon_i(\alpha) - 1/n \sum_{j=1}^n \epsilon_j(\alpha)$. (The centering ensures that $\hat{F}_n(\alpha)$ has mean 0, like F.)

Bootstrap observations are obtained by adding bootstrap errors $\epsilon_1^*, \ldots, \epsilon_n^*$, i.i.d. $\hat{F}_n(\alpha)$ to the pilot curve $\hat{\mu}_\alpha(x_i)$, i.e.,

$$y_i^* = \hat{\mu}_\alpha(x_i) + \epsilon_i^*, \qquad i = 1, \ldots, n. \tag{16}$$

The bootstrap nonparametric estimator of the curve $\hat{\mu}_\alpha(x)$ with bandwidth λ is

$$\hat{\mu}_\lambda^*(x) = \frac{1}{n\lambda} \sum_{j=1}^n K\left(\frac{x - x_j}{\lambda}\right) y_j^*. \tag{17}$$

So, an estimator of the (weighted) estimation risk $R_F(\mu, \hat{\mu}_\lambda) = 1/n \sum_{i=1}^n E(w_i L[\mu(x_i), \hat{\mu}_\lambda(x_i)])$ is $R_{\hat{F}_n(\alpha)}(\hat{\mu}_\alpha, \hat{\mu}_\lambda^*) = 1/n \sum_{i=1}^n E^*_{\hat{F}_n(\alpha)}(w_i L[\hat{\mu}_\alpha(x_i), \hat{\mu}_\lambda^*(x_i)])$.

This approach was taken, for instance by Faraway (1990), who used the bandwidth minimizing least squares CV for the pilot. He showed that the bootstrap difference $n^{2/5}[\hat{\mu}_\lambda^*(x) - \hat{\mu}_\alpha(x)]$, for $\lambda = cn^{-1/5}$ has the same asymptotic distribution as $n^{2/5}[\hat{\mu}_\lambda(x) - \mu(x)]$, provided that $n\alpha^5 \to \infty$ while $\alpha \to 0$ and $n \to \infty$. This condition implies use of an oversmoothed pilot curve and is necessary to take care of the bias of $\hat{\mu}_\lambda(x)$ in estimating $\mu(x)$. (Note that the bandwidth minimized by CV is $O(n^{-1/5})$ and *does not* satisfy the condition.) Härdle and Bowman (1988)

used a pilot bandwidth $\alpha = O(n^{-1/5})$, i.e., of the same order as λ, and therefore explicitly corrected for the bias by estimating it. This involves estimating the second derivative $\mu''(x)$. Hall (1990), using a moving average, took a different approach which however also resulted in using an oversmoothed pilot estimator.

Example 3 (continued)

To estimate the integrated squared error estimation risk $R_2(\mu^{(p)}, \hat{\mu}_\lambda^{(p)})$, one can use the derivatives of $\hat{\mu}_\alpha(x)$ and $\hat{\mu}_\lambda^*(x)$. However the conditions on the pilot bandwidth necessary to obtain a consistent estimator of the best λ for the derivative problem require further study.

Example 4 (continued)

In density estimation, the estimation risk is $R(f, \hat{f}_\lambda) = E_f L(f, \hat{f}_\lambda)$ where the observations used in computing $\hat{f}_\lambda$ are i.i.d. from the density f. To use the bootstrap in this case, all we need is an estimator of f. Resampling from the empirical distribution function $\hat{F}_n$ does not work because it is not differentiable. So as in nonparametric regression, a pilot estimator $\hat{f}_\alpha$ is used and the bootstrap estimator of estimation risk is $R_{\hat{f}_\alpha}(\hat{f}_\alpha, \hat{f}_\lambda^*) = E^*_{\hat{f}_\alpha} L(\hat{f}_\alpha, \hat{f}_\lambda^*)$ (for example see Léger and Romano 1990a). Under regularity conditions and provided that both $n\alpha^5/\log(n) \to \infty$ and $\alpha \to 0$, the bootstrap is consistent in finding the optimal smoothing parameter λ. Again, note that the optimal smoothing parameter for estimating f is $O(n^{-1/5})$ and does not satisfy the condition for the pilot. Instead an oversmoothed estimator $\hat{f}_\alpha$ must be used.

Taylor (1989) used $R_{\hat{f}_\alpha}(\hat{f}_\lambda, \hat{f}_\lambda^*)$, i.e., the same bandwidth λ is used in the pilot estimator. He recognized that this risk estimator has nonnegligible bias and corrected it through a leave-one-out estimator in the spirit of Section 4. So the corrected risk estimator is a combination of a bootstrap method and a leave-one-out method. Faraway and Jhun (1990) use a method similar to Léger and Romano (1990a), choosing the pilot to minimize (13). Although this choice does not satisfy the asymptotic condition of Léger and Romano their simulations demonstrate good results. As in nonparametric regression, Hall (1990) uses a smaller bootstrap sample size $n_1 \ll n$ from the empirical distribution function $\hat{F}_n$ and also a smaller bootstrap bandwidth $\lambda_1 = O(n_1^{-1/5})$ while f is replaced by $\hat{f}_\lambda$.

These examples show that the bootstrap can be used for choosing bandwidths using pre-

diction or estimation risk from different loss functions, as needed. But care must be exercised in applying a bootstrap algorithm. For instance, the pilot curve must be oversmoothed, which means that the pilot bandwidth should not be chosen by cross-validation. Many practical questions remain, including robustness to the choice of the pilot bandwidth.

On the other hand, the bootstrap approach provides two major advantages over cross-validation: it can directly approximate estimation risk, and the bootstrap observations can also be used to construct approximate pointwise and simultaneous confidence bands, e.g., Härdle and Bowman (1988), Härdle and Marron (1991) and Faraway (1990). Paraphrasing Faraway (1990), "construction of confidence bands is the major advantage of the bootstrap."

6 Conclusion

In the last 20 years, many nonparametric statistical procedures have been introduced and much research has been devoted to the crucial problem of choosing tuning parameters from data. In this paper, we have tried to study the problem of choosing tuning parameters by outlining some general principles and by giving examples illustrating both successes and failures.

Cross-validation estimates prediction risk through averaging prediction losses. It is simple to compute and does not require knowing or even estimating the "true" model. The method is automatic in that it can easily be adapted to problems that involve prediction. However, CV approximates risk by average loss, and this can lead to inconsistent estimators if terms with expectation zero converge too slowly. Such problems do not arise with the bootstrap estimator as it is a bona fide expectation. Also CV cannot be used in problems where only estimation risk is suitable.

Bootstrap estimators of risk require estimating a model. When the model is well defined, only the distribution of the observations needs to be approximated. In such cases, simulation from the empirical distribution function $\hat{F}_n$ or a smoother estimator of the distribution is usually adequate. In some problems, the model itself must be estimated, leading to a number of possible bootstrap methods. An important advantage of the bootstrap over cross-validation is that both estimation and predictions risks can be estimated through an appropriate bootstrap algorithm. The bootstrap algorithm also provides tools to set approximate confidence bands around the

estimated curves or predicted values, an important advantage over competing methods.

When the model must be estimated, the bootstrap is not fully automatic. Often a pilot estimator of the model is required. This is also true of a number of asymptotic methods which require, for instance, the estimation of the integrated squared second derivative, as a preliminary to the setting of the appropriate bandwidth. Bootstrap methods can be adapted to many different contexts, but a theoretical study is often required to make sure that the adaptation is adequate. Thus they may be more difficult to use than CV estimators. Finally, bootstrap estimators are usually computer intensive.

Another class of methods, which have not been discussed here are the so-called plug-in estimators (e.g. Park and Marron, 1990). They rely on computing the leading terms of (asymptotic) risk based on the known F and then estimating unknown terms from the data. When the unknown terms are functions of F, they may be estimated by bootstrap methods, leading to a bootstrap estimator of the *asymptotic risk*, whereas the bootstrap methods of the previous section estimate *finite sample risk*. However, plug-in methods in general need not use bootstrap estimation. They require delicate asymptotic computations which are problem and loss dependent.

A number of risk estimators have been developed based on the leave-some-out principle. Although these are often called cross-validated risks, they are not actually based on model validation. In most examples, the risk $R(\theta, \hat{\theta}_\lambda)$ is approximated by an average of losses of the form $L(\tilde{\theta}, \hat{\theta}_\lambda)$, where $\tilde{\theta}$ is another estimator of θ not depending on λ. To reduce biases introduced by computing both estimators on the same data, the data is split — one part is used for computing $\tilde{\theta}$ and the other for computing $\hat{\theta}_\lambda$ and the average is taken over appropriate sets of splits. An example is the divided difference estimator for kernel regression derivatives. Leave-some-out estimators that work in one context often do not work as well in related problems. They have to be tailored to the loss function and to the problem. One such example is the so-called LSCV estimator for kernel density estimation which uses the fact that the loss is L_2 and that a certain term to be estimated can be written as an expectation. This method cannot be extended to estimation of the integral or derivative of the density.

Modified maximum likelihood is another leave-one-out method used to choose tuning pa-

rameters. It is based on the idea of maximizing the "likelihood" of leave-one-out estimators evaluated at the deleted points. While the heuristics of the method are appealing, it need not lead to a good choice of tuning parameter. For instance, in density estimation, the method leads to the optimization of the Kullback-Leibler risk, and is known not to be consistent for long-tailed distributions (Schuster and Gregory, 1981) whereas in binary regression, it leads to the optimization of an ad hoc risk function.

Numerical comparisons of the different methods is beyond the scope of this paper. Nevertheless, we hope that this paper will be a useful guide in helping to develop new methods to choose tuning parameters.

References

Altman, N. and Léger, C. (1994a). On the optimality of prediction based selection criteria and the convergence rates of estimators. Rapport de recherche STT 94–1, Dép. de mathématiques et de statistique, Univ. de Montréal.

Altman, N. and Léger, C. (1994b). Bandwidth selection for kernel distribution function estimation. *Journal of Statistical Planning and Inference* (in press).

Altman, N. S. and MacGibbon, B. (1993). Consistent bandwidth selection for kernel binary regression. Biometrics Unit Tech. report BU-1126-M, Cornell Univ.

Azzalini, A., Bowman, A. W. and Härdle, W. (1989). On the use of nonparametric regression for model checking. *Biometrika* **76** 1–11.

Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika* **71** 353–360.

Breiman, L., Friedman, J. H., Olshen, R. and Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth, Belmont, CA.

Clark, R. M. (1975). A calibration curve for radiocarbon dates. *Antiquity* **49** 251–266.

Duin, R. P. W. (1976). On the choice of smoothing parameters for Parzen estimators of probability density functions. *IEEE Trans. Comput.* **C-25** 1175–1179.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Ann. Statist.* **7** 1–26.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap.* Chapman and Hall, London.

Eubank, R. L. (1988). *Spline Smoothing and Nonparametric Regression.* Dekker, New York.

Faraway, J. J. (1990). Bootstrap selection of bandwidth and confidence bands for nonparametric regression. *J. Statist. Comput. Simul.* **37** 37–44.

Faraway, J. J. and Jhun M. (1990). Bootstrap choice of bandwidth for density estimation. *J. Amer. Statist. Assoc.* **85** 1119–1122.

Friedman, J. H. (1991). Multivariate adaptive regression splines (with discussion). *Ann. Statist.* **19** 1–141.

Gasser, T. and Müller, H.-G. (1984). Estimating regression functions and their derivatives by the kernel method. *Scand. J. of Statist.* **11** 171-185.

Geisser, S. (1975). The predictive sample reuse method with applications. *J. Amer. Statist. Assoc.* **70** 320–328.

Habbema, J. D. F., Hermans, J. and van den Broek, K. (1974). A stepwise discriminant analysis program using density estimation. In *Compstat 1974, Proceedings in computational statistics* (G. Bruckmann) 101–110. Physica Verlag, Wien.

Hall, P. (1983). Large sample optimality of least squares cross-validation in density estimation. *Ann. Statist.* **11** 1156–1174.

Hall, P. (1990). Using the bootstrap to estimate mean squared error and select smoothing parameter in nonparametric problems. *J. Multivariate Anal.* **32** 177–203.

Hall, P. and Martin, M. A. (1988). On bootstrap resampling and iteration. *Biometrika* **75** 661–671.

Härdle, W., (1990). *Applied Nonparametric Regression*, Cambridge Univ. Press.

Härdle, W. and Bowman, A. W. (1988). Bootstrapping in nonparametric regression: Local adaptive smoothing and confidence bands. *J. Amer. Statist. Assoc.* **83** 102–110.

Härdle, W., Hall, P. and Marron, J. S. (1988). How far are automatically chosen regression smoothing parameters from their optimum? *J. Amer. Statist. Assoc.* **83** 86–95.

Härdle, W. and Marron, J. S. (1985). Optimal bandwidth selection in nonparametric regression function estimation. *Ann. Statist.* **13** 1465–1481.

Härdle, W. and Marron, J. S. (1991). Bootstrap simultaneous error bars for nonparametric regression. *Ann. Statist.* **19** 778–796.

Hastie, T. J. and Tibshirani, R. J. (1990) *Generalized Linear Models.* Chapman and Hall, London.

Léger, C. and Romano, J. P. (1990a). Bootstrap choice of tuning parameters. *Ann. Inst. Statist. Math.* **42** 709–735.

Léger, C. and Romano, J. P. (1990b). Bootstrap adaptive estimation: The trimmed-mean example. *Canad. J. Statist.* **4** 297–314.

Léger, C., Politis, D. N. and Romano, J. P. (1992). Bootstrap technology and applications. *Technometrics* **34** 378–398.

Müller, H.-G., Stadtmüller, U. and Schmitt, T. (1987). Bandwidth choice and confidence intervals for derivatives of noisy data. *Biometrika* **74** 743–749.

Nadaraya, E. A. (1964). On estimating regression. *Theory Probab. Appl.* **9** 141–142.

Park, B. U. and Marron, J. S. (1990). Comparison of data-driven bandwidth selectors *J. Amer. Statist. Assoc.* **85** 66–72.

Priestley, M. B. and Chao, M. T. (1972). Non-parametric function fitting. *J. Roy. Statist. Soc. Ser. B* **34**, 385–392.

Pruitt, R. C. (1988). Cross-validation in the one sample location problem. Tech. report No. 510, School of Statistics, Univ. of Minnesota.

Rosenberger, J. L. and Gasko, M. (1983). Comparing location estimators: Trimmed means, medians, and trimean. In *Understanding Robust and Exploratory Data Analysis* (D. C. Hoaglin, F. Mosteller and J. W. Tukey, eds.) 297–338. Wiley, New York.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.* **27** 832–837.

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scand. J. Statist.* **9** 65–78.

Sarda, P. (1993). Smoothing parameter selection for smooth distribution functions. *J. Statist. Plann. Inference* **35** 65–75.

Schuster, E. F. and Gregory, G. G. (1981). On the nonconsistency of maximum likelihood

nonparametric density estimators. In *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface* (W. F. Eddy, ed.) 295–298. Springer, Berlin.

Shao, J. (1993). Linear model selection by cross-validation. *J. Amer. Statist. Assoc.* **88** 486–494.

Shao, J. and Wu, C. F. J. (1989). A general theory for jackknife variance estimation. *Ann. Statist.* **17** 1176–1197.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, London.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. Ser. B* **36** 111–147.

Stone, M. (1977). Asymptotics for and against cross-validation. *Biometrika* **64** 29–35.

Stone, C. J. (1984). An asymptotically optimal window selection rule for kernel density estimates. *Ann. Statist.* **12** 1285–1297.

Taylor, C. C. (1989). Bootstrap choice of the smoothing parameter in kernel density estimation. *Biometrika* **76** 705–712.

Wahba, G. and Wold, S. (1975). A completely automatic French curve: Fitting spline functions by cross validation. *Comm. Statist. A—Theory Methods* 4 1–17.